LA SIMULATION DE MODÈLE SUR ORDINATEUR EN TANT QUE MÉTHODE DE RECHERCHE: LE CAS CONCRET DE L'ÉTUDE DE LA DISTRIBUTION ÉCHANTILLONNALE DES ESTIMATEURS EN TESTING ADAPTATIF EN FONCTION DES RÈGLES D'ARRÊT

Gilles Raîche
Conseiller pédagogique - Cégep Joliette - De Lanaudière

PROLOGUE

Effectuer une simulation sur ordinateur offre l'avantage d'éviter les aléas des recherches dont les observations proviennent de sujets humains. Lorsque l'on travaille avec des sujets humains des problèmes qui ne sont pas toujours prévisibles peuvent se produire. Des phénomènes d'attrition ainsi que des contraintes à la sélection des sujets et à l'équivalence des groupes de comparaison ne peuvent pas toujours être prévus et contrôlés.

Une simulation sur ordinateur permet ce contrôle. Cependant d'autres aléas peuvent se produire. Le présent texte illustre cette situation à partir du cas concret de la simulation de la modélisation de l'évaluation adaptative des apprentissages assistée par ordinateur (*computer adaptive testing, CAT*). Des problèmes rencontrés en cours de route lors de l'étude de la distribution échantillonnale des estimateurs en testing adaptatif seront présentés.

La prochaine section illustre le cas concret et l'objet de recherche à analyser. Une introduction décrit ce qu'est un test adaptatif. Ensuite la théorie des réponses aux items est présentée comme outil mathématique utile en testing adaptatif. La structure d'un test adaptatif est décrite. Enfin, la méthodologie de recherche visant à permettre l'étude de la distribution échantillonnale des estimateurs est présentée. Une dernière section, un épilogue, termine le texte. Les aléas rencontrés en cours de réalisation de l'étude y sont décrits.

INTRODUCTION

Lorsqu'un éducateur physique désire comparer la performance de ses étudiantes et de ses étudiants au saut en hauteur il peut adopter deux stratégies.

Dans la première, il place une barre horizontale à différentes hauteurs prédéterminées qui seront les mêmes pour tous. Il demande ensuite à chacun et chacune de tenter un saut à la barre aux différentes hauteurs. Le nombre de sauts réussis servira à comparer les étudiants entre eux.

Il peut aussi adopter une stratégie de comparaison où la hauteur des sauts à effectuer est ajustée en fonction de l'habileté de chacun des étudiants. Au premier saut, la barre est placée à une hauteur jugée moyenne. Si l'étudiant réussit le saut, la hauteur de la barre est augmentée. Ceci jusqu'au moment où l'étudiant rate son saut. Si, au contraire, l'étudiant ne réussit pas son saut à la première tentative, la hauteur de la barre est diminuée jusqu'à ce qu'il réussisse. La stratégie de comparaison est adaptée à l'habileté de l'étudiant. Bien sûr, le nombre de sauts réussis ne peut plus servir à comparer les étudiants entre eux puisque la hauteur des sauts varie d'un étudiant à un autre. La hauteur du dernier saut réussi serait une mesure plus réaliste et permettrait de connaître de façon plus précise la performance de chacun de ces étudiants.

Lorsqu'un enseignant désire comparer l'habileté en mathématique de ses étudiants il peut leur administrer un examen à choix multiples qui est le même pour tous. Le nombre de réponses réussies devient l'unité de comparaison entre les étudiants. Cette situation est équivalente à la première stratégie adoptée par l'éducateur physique. Plus le nombre de bonnes réponses, ou de sauts réussis, est élevé plus la performance de l'étudiant est élevée. Dans les deux situations il n'est aucunement nécessaire de connaître la hauteur du saut ou l'habileté réelle en mathématique. Seul le nombre de réussites est suffisant.

L'enseignant pourrait aussi adopter une stratégie ajustée au niveau d'habileté en mathématique de chacun des étudiants. Cette méthode est similaire à la seconde stratégie utilisée par l'éducateur physique. Un premier item à choix multiples et de niveau d'habileté moyen est présenté à l'étudiant. Si l'étudiant le réussi, un second item de niveau d'habileté supérieur lui est administré. Et ainsi de suite. A l'inverse, lorsque l'étudiant rate un item, un item de niveau d'habileté inférieur lui est administré. Comme dans le cas de la mesure de la performance en saut en hauteur, le nombre de bonnes réponses ne peut plus servir de mesure du niveau d'habileté en mathématique de l'étudiant. Le nombre de questions administrées et le niveau d'habileté en mathématique mesuré varient d'un étudiant à un autre. L'examen, ou le test, est dit adaptatif.

Contrairement au cas du saut en hauteur il est toutefois impossible de mesurer directement le niveau d'habileté en mathématique. Comme dans un rêve, le nombre de bonnes réponses n'est que le contenu manifeste. La vraie signification de ce contenu, le niveau d'habileté en mathématique, n'est pas directement visible. C'est le contenu latent, le trait latent.

Des modèles probabilistes ont été proposés pour estimer de telles habiletés cognitives. La théorie des réponses aux items (TRI) est un de ces modèles. Quelles sont les caractéristiques des estimateurs des habiletés calculés à partir de la théorie des réponses aux items? Plus précisément, puisque ce sont des estimateurs probabilistes, quelle est la distribution échantillonnale des estimateurs en testing adaptatif en fonction des règles d'arrêt? Le présent projet désire tenter de répondre à cette question.

La prochaine section présente la théorie des réponses aux items. Ensuite les algorithmes courants en testing adaptatif sont exposés en mettant en relief les diverses règles d'arrêt. Enfin une méthodologie est proposée pour étudier la distribution échantillonnale des estimateurs en testing adaptatif en fonction des règles d'arrêt.

LA THÉORIE DES RÉPONSES AUX ITEMS

La théorie des réponses aux items a été proposée pour répondre aux lacunes de la théorie classique des tests basée sur le nombre de bonnes réponses (Humbleton, Swaminathan et Rogers, 1990). La théorie des réponses aux items permet, contrairement à la théorie classique, d'utiliser des paramètres associés aux items qui sont indépendants du groupe particulier à l'intérieur duquel ils ont été obtenus. Les paramètres sont dits invariants par rapport au groupe. La proportion de réussite à un item, utilisée en théorie classique, n'est pas invariante par rapport au groupe où elle est estimée.

De plus, dans la théorie des réponses aux items, le niveau d'habileté estimé n'est pas dépendant des items utilisés pour effectuer la mesure. Il est question d'invariance par rapport aux items. Cette invariance du trait par rapport aux items est une caractéristique intéressante en testing adaptatif où justement les items sélectionnés sont différents d'un individu à un autre. C'est pourquoi la théorie des réponses aux items est utilisée en testing adaptatif.

Plusieurs modélisations de la théorie des réponses aux items ont été proposées. Certaines de ces modélisations dépendent du type de réponse aux items: réponses dichotomiques (Lord, 1952, 1980; Lord et Novick, 1968), réponses polychotomiques (Baker, 1992), réponses ordonnées (Samejima, 1977), réponses polychotomiques partiellement ordonnées (Wilson, 1992) et réponses continues (Samejima, 1973). D'autres modélisations dépendent du type de trait postulé: trait catégoriel (classes latentes), trait continu (trait latent) ou trait hybride (classes latentes et trait latent). Dans d'autres cas l'estimation du trait est multidimensionnelle (Goldstein et Wood, 1989; McDonald, 1982). Des modèles non-paramétriques existent aussi (Ramsey, 1991, 1993).

La famille de modèles retenue ici est paramétrique, le type de réponses aux items est dichotomique (bonne ou mauvaise réponse) et le trait latent est unidimensionnel. À l'intérieur de cette famille trois modèles sont utilisés en testing adaptatif. Ce sont les modèles logistiques à un, deux et trois paramètres. Ces trois modèles sont présentés. La section se termine sur le choix d'un de ces modèles et la justification de ce choix.

Le modèle à un paramètre

Le modèle logistique à un paramètre postule que la probabilité de réussite d'un item est fonction du niveau d'habileté (θ) et du niveau de difficulté de l'item (b).

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}} \qquad (1)$$

La figure 1 présente trois courbes qui se distinguent uniquement par le niveau de difficulté de l'item correspondant. Ce sont les courbes caractéristiques des items (CCI). Comme dans la plupart des modèles de réponses aux items, le niveau d'habileté est ramené à une moyenne de 0 et à une variance unitaire. Cette standardisation permet une interprétation du niveau d'habileté en fonction de la loi normale centrée réduite.

Le modèle suppose, comme l'indiquent la fonction 1 et la figure 1, que lorsque le niveau d'habileté est très faible la probabilité de réussite de l'item est nulle. Au contraire, plus le niveau d'habileté est élevé plus la probabilité de réussite de l'item se rapproche de 1. Le niveau de difficulté de l'item (b) correspond au niveau d'habileté où l'item permet la plus grande discrimination et où l'erreur de mesure est la plus faible. En testing adaptatif, l'objectif est d'administrer des items dont le niveau de difficulté permet la meilleure discrimination du niveau d'habileté.

Dans une situation de testing adaptatif plusieurs items de différents niveaux de difficulté sont administrés. Le niveau d'habileté mesuré par une séquence de réponses à des items peut être estimé en maximisant la probabilité du niveau d'habileté déterminé par les courbes caractéristiques des items. La fonction 2 est donc maximisée.

$$P(x_i \mid \theta_i, \beta) = \prod_j P_j(\theta_i)^{x_{ij}} Q_j(\theta_i)^{1-x_{ij}} \qquad (2)$$

La figure 2 montre les niveaux d'habileté qui maximisent trois patrons de réponses différents aux trois items de la figure 1. Ainsi, lorsque les deux premiers items (b= -1 et 0) sont ratés tandis que le troisième item

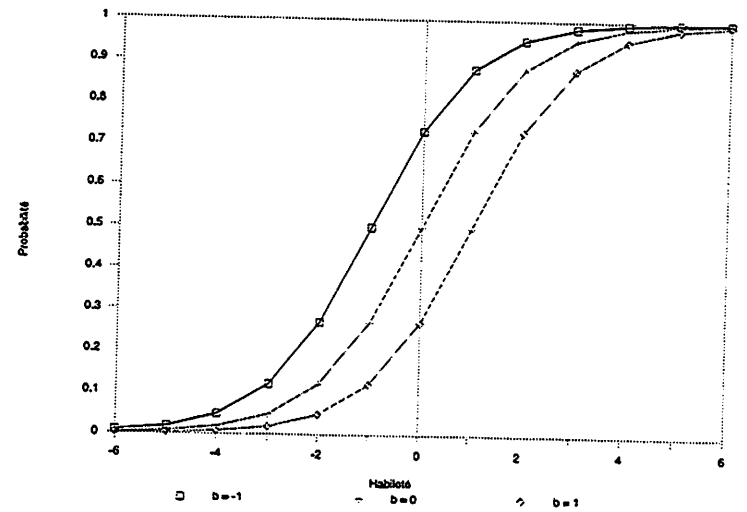

Figure 1
Courbe caractéristique d'item (CCI) du modèle à un paramètre selon trois niveaux de difficulté

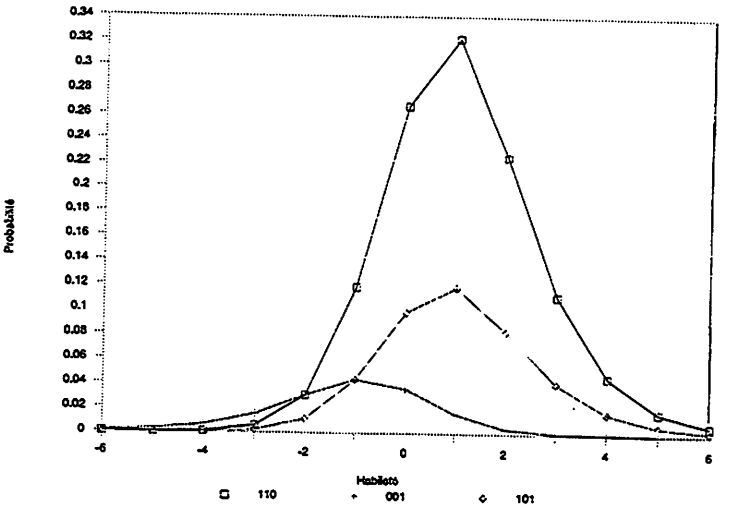

Figure 2
Estimation de l'habileté dans le modèle à un paramètre selon trois configurations de réponse à trois items

(b = 1) est réussi, le niveau d'habileté qui maximise la probabilité d'apparition de ce patron de réponse est égal à environ -1. Dans ce cas le niveau d'habileté est relativement faible.

Le modèle logistique à deux paramètres

Le modèle logistique à deux paramètres propose l'addition, au niveau de difficulté (b), d'un second paramètre: la discrimination (a). Selon ce modèle, les items ne partagent pas tous le même pouvoir de discrimination lorsque le niveau d'habileté est égal au niveau de difficulté. De façon algébrique, l'indice de discrimination correspond à la pente de la courbe caractéristique de l'item lorsque le niveau d'habileté est égal au niveau de difficulté de l'item.

$$P(\theta) = 1 + \frac{1}{1 + e^{-a(\theta - b)}} \qquad (3)$$

La figure 3 présente trois courbes caractéristiques d'item lorsque le niveau de difficulté (b) est maintenu à 0 tandis que l'indice de discrimination (a) varie de .5 à 2. Il est clairement visible que les courbes caractéristiques des items n'affichent pas la même pente lorsque le niveau d'habileté est égal à 0. La figure 4 montre les niveaux d'habileté qui maximisent les mêmes patrons

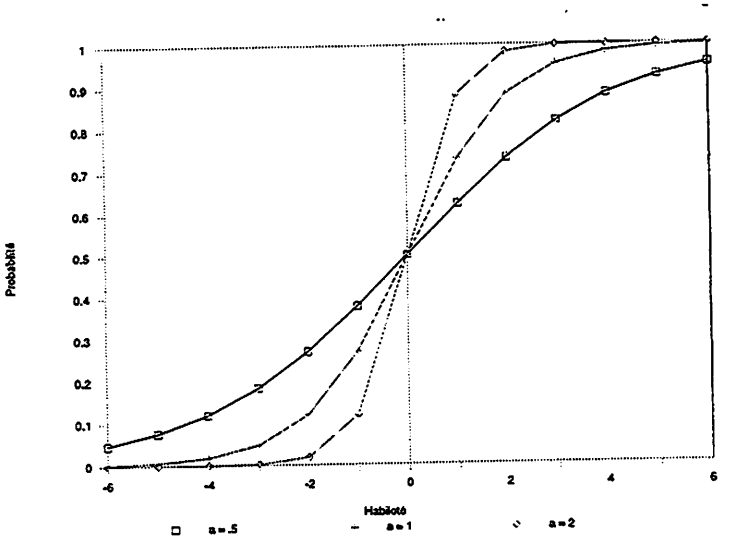

Figure 3
Courbe caractéristique d'item (CCI) du modèle à deux paramètres selon trois niveaux de discrimination

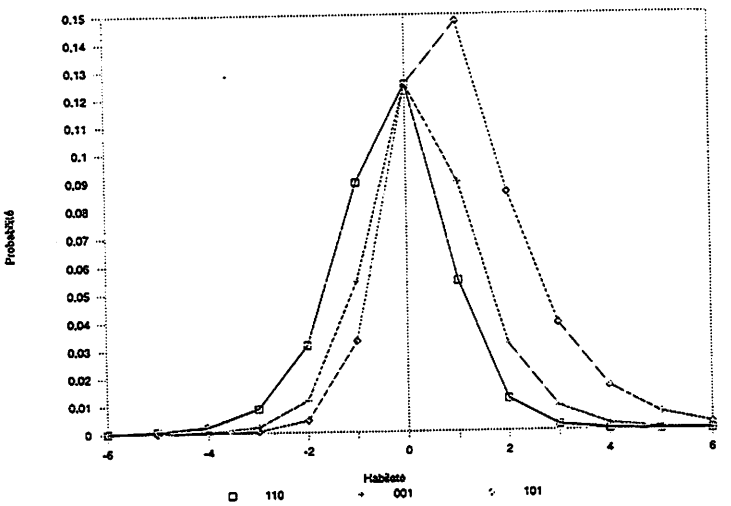

Figure 4
Estimation de l'habileté dans le modèle à trois paramètres selon trois configurations de réponse à trois items

de réponses présentés à la figure 2. Les paramètres d'item sont cependant ceux utilisés pour créer la figure 3. La probabilité du patron de réponse 001 est maximisée lorsque le niveau d'habileté est égal à 0.

Le modèle logistique à trois paramètres

Le modèle logistique à trois paramètres ajoute un troisième élément de paramétrisation aux deux modèles précédents: l'indice de pseudo-chance (c) (*pseudo-guessing*). Selon ce modèle, la probabilité de réussir un item n'est pas nécessairement nulle lorsque le niveau d'habileté est très faible. Des facteurs externes au niveau d'habileté peuvent affecter la probabilité de réussite. Par exemple, dans un choix de réponse *vrai ou faux*, la réponse *oui* pourrait être naturellement préférée par les individus dont le niveau d'habileté est très faible. La figure 5 montre de tel type d'items. L'item dont le paramètre de pseudo-chance (c) est égal à .8 est réussi huit fois sur dix lorsque le niveau d'habileté est très faible. La figure 6 présente comment sont maximisées les probabilités d'apparition des mêmes patrons de réponses suggérés aux figure 2 et 4. Il est à remarquer que le patron de réponses 001 est maximisé lorsque le niveau d'habileté est inférieur à -6. L'équation du modèle logistique à trois paramètres est la suivante:

$$P(\theta) = c + \frac{1 - c}{1 + e^{-a(\theta - b)}} \qquad (4)$$

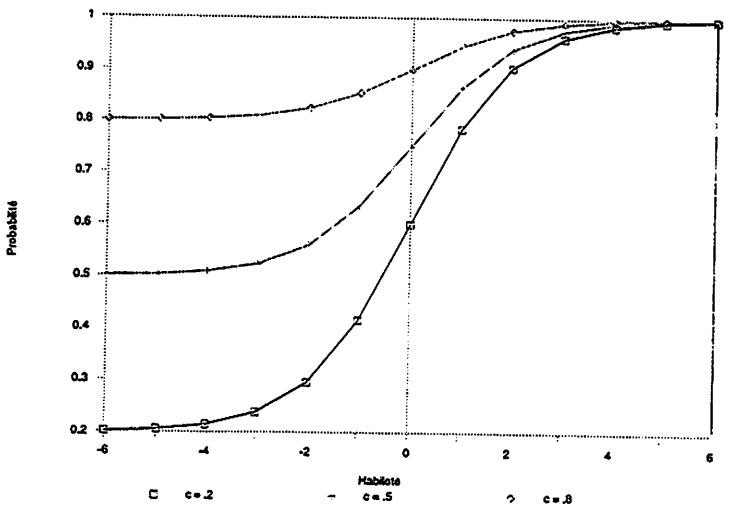

Figure 5
Courbe caractéristique d'item (CCI) du modèle à trois paramètres selon trois niveaux de pseudo-chance

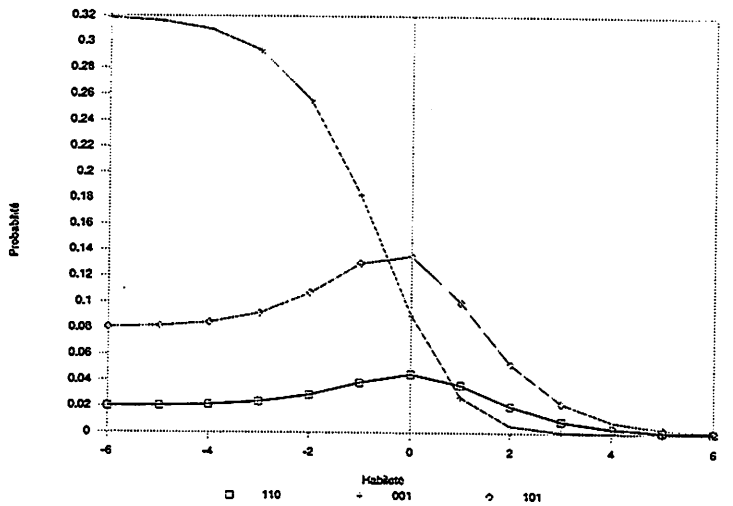

Figure 6
Estimation de l'habileté dans le modèle à deux paramètres selon trois configurations de réponse à trois items

Choix d'un modèle

Les modèles logistiques les plus utilisés sont les modèles à un et à trois paramètres. Le premier permet l'utilisation d'algorithmes d'estimation du niveau d'habileté rapides et efficaces. Le second, à trois paramètres, est plus lourd à utiliser et risque, quoique rarement, de ne pas fournir de solution unique. L'estimation est cependant plus précise.

Le modèle à trois paramètres présente, selon Wainer (1990), un inconvénient potentiel en testing adaptatif. La plupart des algorithmes de sélection d'items ont pour objectif de maximiser la discrimination au niveau d'habiletés provisoires estimé. En recherchant cette maximisation les items dont l'indice de discrimination est élevé sont sélectionnés plus fréquemment. Selon Wainer, cette situation risque de provoquer l'apparition d'items dont les contenus sont très similaires au détriment d'autres contenus. Cet inconvénient potentiel n'existe pas dans les situations de testing standards où l'équilibre des contenus est facilement sous contrôle. C'est pourquoi le présent projet adopte le modèle logistique à un paramètre où cet inconvénient est absent.

La prochaine section présente la structure d'un test adaptatif, sa mécanique.

STRUCTURE D'UN TEST ADAPTATIF

Dans un test adaptatif où l'on essaie de présenter des items se rapprochant le plus possible du niveau réel d'habileté, des décisions doivent être prises en ce qui concerne les caractéristiques du ou des premiers items administrés. C'est la règle de départ. Suite à la performance à ou aux premiers items, d'autres items se rapprochant de plus en plus du niveau réel d'habileté sont proposés. C'est la règle de sélection. Enfin, un ou des critères ayant pour but de décider de mettre fin à la situation de mesure doivent être adoptés. Il est question de la règle d'arrêt. La figure 7 montre la structure d'un test adaptatif. Pour chacune des règles, des stratégies proposées par la littérature sont présentées. Par choix, l'approche par mini-tests (*testlets*) de Wainer (1990) n'est pas présentée.

Règle de départ

Un test adaptatif doit débuter de quelque façon. Une règle de départ basée sur l'information à priori disponible au sujet du niveau d'habileté doit être adoptée. Généralement, la moyenne de la population d'où est tiré l'individu en situation de testing est une estimation provisoire de départ raisonnable du niveau d'habileté (Wainer, 1990). Le premier item présenté est donc un item dont les paramètres permettent une discrimination optimale lorsque le niveau d'habileté est égal à la moyenne de la population. Dans certain

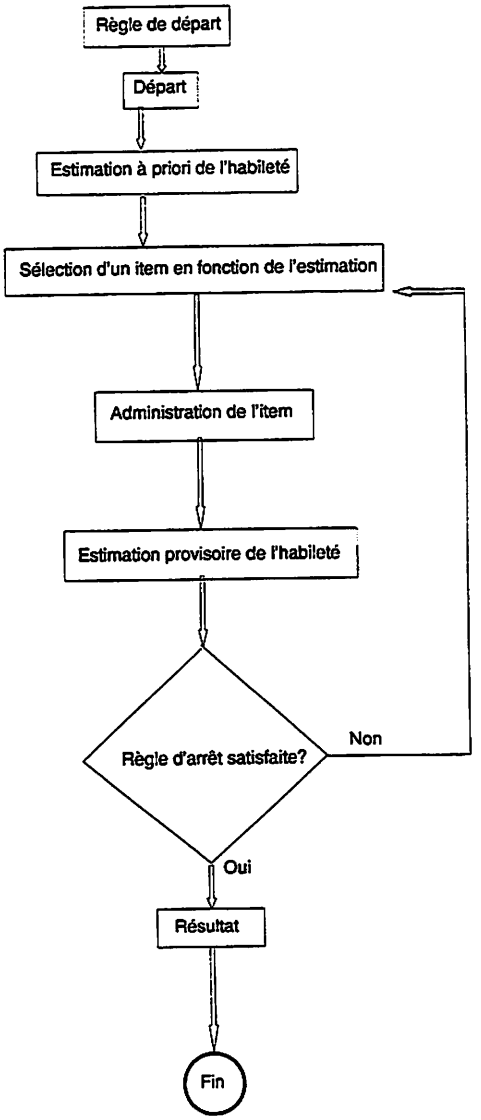

Figure 7
Structure d'un test adaptatif

cas l'estimation provisoire à priori peut être améliorée en utilisant de l'information supplémentaire comme le sexe, l'âge ou même une appréciation subjective de la part de l'individu testé.

Règle de sélection

Selon la performance au premier item, un second item optimal doit être sélectionné et ensuite administré. Et ainsi de suite jusqu'à ce que la règle d'arrêt soit satisfaite. Deux règles de sélection du prochain item sont habituellement en usage.

La première de ces règles consiste à obtenir une estimation provisoire du niveau d'habileté par la méthode de vraisemblance maximale (*maximum likelihood*). Ensuite un item dont les caractéristiques permettent une estimation optimale du niveau d'habileté est administré. C'est la méthode suggérée par Lord (1980) qui consiste à maximiser la fonction suivante:

$$P(x_i|\theta_i,\beta) = \prod_j P_j(\theta_i)^{x_{ij}} Q_j(\theta_i)^{1-x_{ij}} \qquad (5)$$

La méthode de vraisemblance maximale a toutefois l'inconvénient de ne pas offrir de solution lorsque les réponses aux items sont soit, toutes des réussites, soit, toutes des échecs. C'est le cas, entre autre, de l'estimation provisoire du niveau d'habileté après l'administration du premier item. C'est pourquoi Wainer (1990) propose une méthode d'estimation bayésienne qui utilise de l'information connue à priori. L'estimation provisoire obtenue devient ainsi une estimation à posteriori. Cette méthode consiste à maximiser la probabilité du niveau d'habileté en incorporant à la fonction 5 la fonction de probabilité à priori du niveau d'habileté.

$$P(x|\theta,\beta) = \prod_j P_j(\theta)^{x_j} Q_j(\theta)^{1-x_j} * P(\theta|y) \qquad (6)$$

En général, il est postulé que la probabilité à priori du niveau d'habileté suit une loi normale centrée réduite. La méthode de vraisemblance maximale n'est en fait qu'un cas particulier de la méthode bayésienne où la loi de probabilité à priori postulée est une loi uniforme.

Règle d'arrêt

Puisque toute bonne chose a une fin, un test adaptatif doit finalement être interrompu. Deux règles d'arrêt sont généralement utilisées. La première consiste à clore la situation de mesure après l'administration d'un nombre fixe et prédéterminé d'items. Aucune règle absolue n'a été arrêtée quant à ce nombre. Selon Wainer (1990), l'administration d'un nombre minimal de 20 items permet d'obtenir une estimation du niveau d'habileté presque identique que ce soit en utilisant une estimation par vraisemblance maximale ou une estimation bayésienne. En fait, dans l'estimation bayésienne, plus le nombre d'items administrés est élevé moins la probabilité à priori a d'impact sur l'estimation.

Une seconde règle d'arrêt utilisée consiste à terminer l'administration du test lorsqu'une précision satisfaisante du niveau d'habileté est obtenue. La précision de l'estimation est mesurée par l'inverse de la fonction d'information:

$$I(\theta) = \sum_j \frac{(P_j')^2}{P_j(\theta)Q_j(\theta)} \qquad (7)$$

$$S(\theta) = \frac{1}{\sqrt{I(\theta)}} \qquad (8)$$

Comme dans le cas de la règle d'arrêt basée sur le nombre d'items administrés, aucune règle absolue n'a été arrêtée quant à la précision à atteindre. La règle d'arrêt basée sur la précision de l'estimation a l'avantage de permettre de fournir une estimation de même qualité à tous les niveaux d'habileté. Ce n'est pas le cas avec la règle d'arrêt basée sur le nombre d'items administrés.

Selon Wainer, le choix d'une règle d'arrêt a un impact sur l'estimation du niveau d'habileté. Le présent projet a comme objectif de vérifier cette hypothèse. Plus précisément, l'impact du nombre d'items fixés à l'avance dans la première règle d'arrêt et du niveau de précision dans la seconde règle sur la distribution échantillonnale de l'estimation du niveau d'habileté sera étudié. Les résultats obtenus devraient permettre d'identifier les conditions d'utilisation optimales de ces règles: nombre minimal d'items à administrer et précision minimale à atteindre.

La prochaine section présente la méthodologie proposée pour vérifier l'hypothèse de recherche.

MÉTHODOLOGIE

Dans le but d'exercer un contrôle strict de la situation de mesure et d'expérimenter sur un nombre important d'unités d'observation, une simulation informatisée est proposée. Les unités d'observation retenues, le déroulement de la simulation, les hypothèses à vérifier et la méthode d'analyse des résultats sont maintenant décrits.

Unités d'observation

La simulation est appliquée à 100 unités d'observation pour chacun des niveaux d'habileté prédéterminés sur

une échelle centrée réduite. La valeur de ces niveaux d'habileté varie entre -3 et 3 par saut de .5. Ces valeurs sont retenues puisque, théoriquement, lorsque le niveau d'habileté suit une loi normale (N(0,1)), elles couvrent la presque totalité des événements possibles. Un saut de .5 semble suffisant pour permettre ensuite des interpolations raisonnables entre ces valeurs.

Déroulement de la simulation

La figure 8 montre le déroulement de la simulation pour la règle d'arrêt basée sur le nombre d'items administrés. Un nombre minimal de 5 items à administrer (n) est fixé au départ. Ensuite, une situation de testing adaptatif est mise en place pour un niveau d'habileté (θ) égal à -3. L'estimation à priori de θ est fixée à la moyenne de la population (0) et une réponse aléatoire est simulée en tenant compte de l'estimation provisoire de θ et du niveau de difficulté (b) associé. La génération d'une réponse aléatoire est effectuée à partir d'une méthode proposée par Cooke, Graven et Clarke (1982). Une nouvelle estimation est effectuée, ainsi que le calcul de la précision de cet estimateur. Si le nombre d'items à administrer n'est pas atteint, une nouvelle réponse aléatoire est simulée. Si le nombre d'items à administrer est atteint, l'information est emmagasinée dans un fichier: θ, estimation de θ , S(θ) et n. Si le nombre de simulations est inférieur à 100, l'estimation à priori de θ est réinitialisée à 0. Sinon, le niveau d'habileté à estimer est augmenté de .5 tant qu'il ne dépasse pas 3. Lorsque cette valeur est dépassée, le nombre d'items à administrer est augmenté de 1. La stratégie recommence tant que le nombre d'items à administrer ne dépasse pas 40. La même stratégie est adoptée pour le déroulement de la simulation lorsque

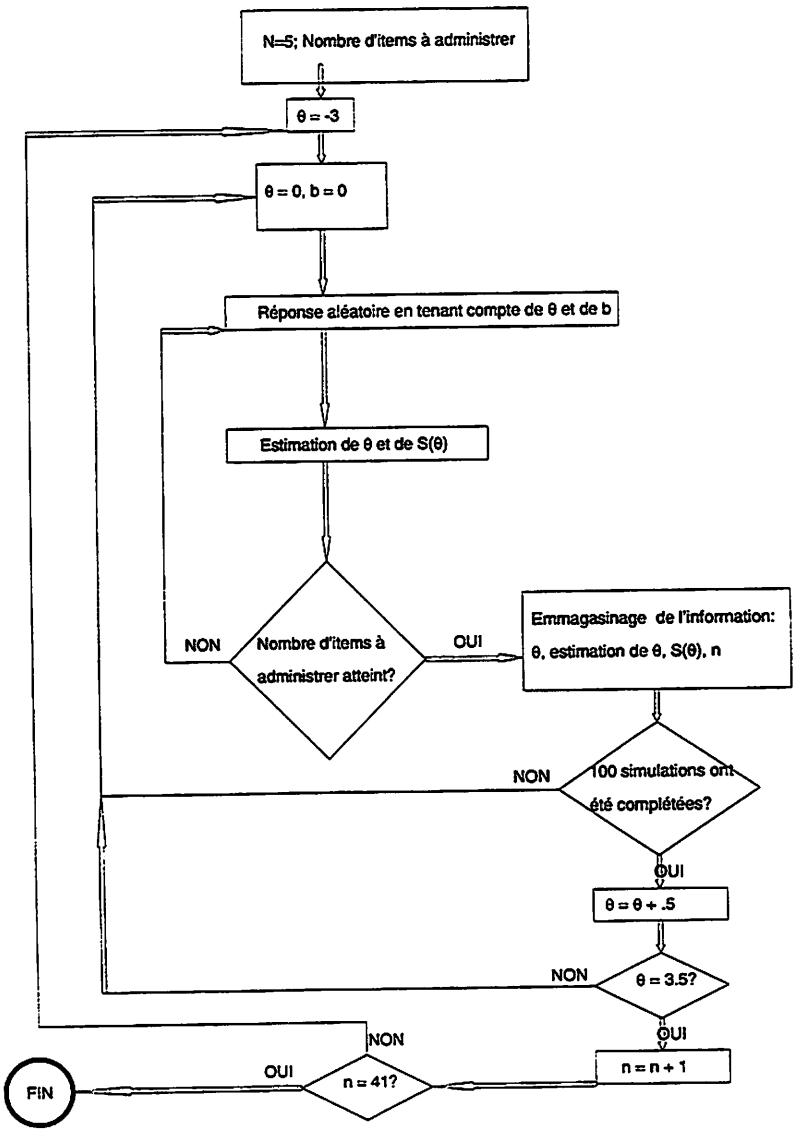

Figure 8
Déroulement de la simulation lorsque la règle est basée sur le nombre d'items administrés

la règle d'arrêt est basée sur la précision de l'estimation.

Hypothèse

La variation de la règle d'arrêt, que ce soit par le nombre d'items à administrer ou par la précision de l'estimateur, affecte la distribution échantillonnale de l'estimateur du niveau d'habileté: forme de la distribution, moyenne et variance.

Méthode d'analyse des résultats

La méthode d'analyse des résultats est basée à la fois sur des représentations graphiques et sur des indices descriptifs des distributions. Ainsi pour chacun des 13 niveaux d'habileté fixés, la distribution échantillonnalle des estimateurs en fonction du nombre d'items à administrer est représentée. Il en est de même en ce qui concerne la précision des estimations à atteindre. Pour chacune des distributions échantillonnales représentée, la moyenne, la variance ainsi que les coefficients d'asymétrie et d'aplatissement sont calculés.

Il est aussi prévu de représenter graphiquement la moyenne de l'estimation, la précision moyenne et l'erreur d'estimation en fonction du nombre d'items administrés. Ceci pour chacun des 13 niveaux d'habileté étudiés. Le nombre moyen d'items administrés, l'erreur moyenne d'estimation et la moyenne de l'estimation seront aussi représentés graphiquement en fonction de la précision à satisfaire.

ÉPILOGUE

Suite à la description du cas concret de l'étude de la distribution des estimateurs en testing adaptatif selon les règles d'arrêt, cette section présente deux aléas rencontrés lors de l'élaboration de cette étude.

Le premier de ces aléas est apparu au moment de programmer le calcul de l'estimation du trait latent. Il s'agissait de maximiser les fonctions 5 (méthode de vraisemblance maximale) et 6 (méthode bayésienne). La littérature ne présente pas toujours les fonctions de probabilité sous-jacentes de la même façon. Ceci est surtout remarquable dans le cas de l'estimation bayésienne où la fonction de probabilité à priori n'est pas toujours explicite: fonction de probabilité ou de probabilité cumulative. De plus les exemples offerts dans la littérature présentent quelquefois des erreurs de calcul. Ceci ne permet pas de se fier aveuglément aux exemples pour vérifier l'exactitude des calculs obtenus par programmation. Il est donc conseillé de consulter plusieurs sources. Enfin le choix du langage de programmation est important. Un langage tout usage tel Fortran, Pascal, Basic ou C est lourd à utiliser. Certains langages sont totalement dédiés à ce type de programmation. Mathcad a été retenu car il permet d'effectuer des calculs symboliques et de manipuler les fonctions sous la même forme graphique que celle qui est présentée dans les textes.

Le second de ces aléas, à propos duquel le chercheur n'a pas encore pris de décision, concerne la stratégie d'échantillonnage des unités d'observation pour chacun des niveaux d'habileté prédéterminés. Est-ce que plus de 100 observations par niveaux seront sélectionnées? Il ne faut pas perdre de vue que si tous les patrons de réponses possibles à 20 items sont retenus 2^{20} (1,048,576) solutions sont possibles. Le nombre risque d'être trop élevé pour permettre le traitement des résultats à partir des logiciels statistiques usuels. De quelle façon doit-on effectuer l'échantillonnage de ces éléments? Un échantillonnage aléatoire ne va pas de soi car les patrons de réponse ne sont pas tous équiprobables. Leur probabilité d'apparition varie en fonction du niveau d'habileté. Ce problème reste donc à résoudre et il est à espérer que sa solution n'entrainera pas l'apparition de nouvelles difficultés.

RÉFÉRENCES

Baker, F.B. (1992). *Item response theory: parameter estimation techniques*. New York, Marcel Dekker.

Cooke, D., Graven, A.H., Clarke, G.M. (1982). *Basic statistical computing*. London, Edward Arnold.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika, 78-1*, 45-51.

Goldstein, H., Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph*, No 7.

Lord, F.M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, Lawrence Erlbaum.

Lord, F.M., Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, Addison-Wesley.

McDonald, R.P. (1982). Linear versus non-linear models in item response theory. *Applied Psychological Measurement, 6*, 379-396.

Ramsey, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56-4*, 611-630.

Ramsey, J.O. (1993). *TESTGRAF: a program for the graphical analysis of multiple choice test and questionnaire data*. Montréal, Département de psychologie, Université McGill.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38-2*, 203-219.

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika, 42-2*, 163-188.

Wainer, H. (1990). *Computerized adaptive testing: a primer*. Hillsdale, Lawrence Erlbaum.

Wilson, M. (1992). The ordered partition model: an extension of the partial credit model. *Applied Psychological Measurement*, 16-4, 309-325. ❖